Quantifying the accessibility of the metagenome by random expression cloning techniques

Gabor, E.M.; Alkema, W.B L; Janssen, D.B.

Published in:
Environmental Microbiology

DOI:
10.1111/j.1462-2920.2004.00640.x

Publication date:
2004

Environmental Microbiology (2004) **6**(9), 879–886 doi:10.1111/j.1462-2920.2004.00640.x

Quantifying the accessibility of the metagenome by random expression cloning techniques

Esther M. Gabor,[1†] Wynand B. L. Alkema[2†] and Dick B. Janssen[1*]

[1]*Department of Biochemistry, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Nijenborgh 4, 9747 AG Groningen, the Netherlands.*

[2]*Center for Genomics and Bioinformatics, Karolinska Institute, Berzelius väg 35, 17177 Stockholm, Sweden.*

Summary

The exploitation of the metagenome for novel biocatalysts by functional screening is determined by the ability to express the respective genes in a surrogate host. The probability of recovering a certain gene thereby depends on its abundance in the environmental DNA used for library construction, the chosen insert size, the length of the target gene, and the presence of expression signals that are functional in the host organism. In this paper, we present a set of formulas that describe the chance of isolating a gene by random expression cloning, taking into account the three different modes of heterologous gene expression: independent expression, expression as a transcriptional fusion and expression as a translational fusion. Genes of the last category are shown to be virtually inaccessible by shotgun cloning because of the low frequency of functional constructs. To evaluate which part of the metagenome might in this way evade exploitation, 32 complete genome sequences of prokaryotic organisms were analysed for the presence of expression signals functional in *E. coli* hosts, using bioinformatics tools. Our study reveals significant differences in the predicted expression modes between distinct taxonomic groups of organisms and suggests that about 40% of the enzymatic activities may be readily recovered by random cloning in *E. coli*.

Introduction

During the past five years, random cloning of microbial DNA directly isolated from environmental materials and subsequent screening of expression libraries for the presence of a desired enzyme activity has become a useful tool for the discovery of novel biocatalysts. The collective genomes of microbes indigenous to a certain habitat, now often referred to as the metagenome (Handelsman *et al*., 1998), are considered to be an almost inexhaustible source of new enzymes (Cowan, 2000). Indeed, screening of the metagenome has already yielded various new biocatalysts (for a recent review see Lorenz and Schleper, 2002), and with steadily improving techniques this number is expected to rise quickly. In most cases, gene banks are screened with activity-based assays as they allow the recovery of completely new types of enzymes without any prior knowledge of the sequence, relying only on the ingenuity of the screening method. Such a functional screening, however, requires gene expression and proper folding of the resulting protein in a heterologous host, most frequently *E. coli*, which is not always easily achieved.

The minimal set of requirements for gene expression includes the presence of a promoter for transcription, and a ribosome binding site (rbs) in the −20 to −1 region upstream of the start codon for initiation of translation. Both sites must be suitable for the expression machinery of the bacterial host cell. Besides these *cis*-acting DNA sequences, the formation of an active protein may also rely on *trans* factors that need to be provided by the host organism such as special transcription factors, inducers, chaperones, cofactors, protein-modifying enzymes, or a proper secretion machinery. Whether or not essential *trans* factors are present in the host is in most cases difficult or even impossible to predict. In contrast, functional *cis* elements can be identified based on DNA sequence analysis (e.g. Gold *et al*., 1981; Staden, 1983; Ermolaeva *et al*., 2000).

To assure the formation of mRNA transcripts of heterologous coding sequences (cds), vectors carrying their own strong promoter (and possibly a transcriptional terminator) are usually employed in expression cloning. In addition, a rbs followed by a bacterial start codon in favourable spacing (9 bp for *E. coli*) is generally supplied close to the multiple cloning site. In such systems, three modes of gene expression can be anticipated: (i) independent gene expression with both the promoter and the rbs provided by the insert (IND); (ii) expression as a transcriptional fusion with only the rbs located on the insert (TRANSC), and (iii) expression as a translational fusion depending on

Received 7 January, 2004; accepted 23 March, 2004. *For correspondence. E-mail D.B.Janssen@chem.rug.nl; Tel. (+31) 50 3634209; Fax: (+31) 50 3634165. †Both authors contributed equally to this work.

both the promoter and the rbs of the vector (DEP) (Fig. 1). Intuitively, it can be understood that the occurrence of a functional translational fusion is very rare and, consequently, the chance of discovering a microbial gene devoid of expression signals that are recognized by its heterologous host is low. Transcriptional fusions, in contrast, are more likely to occur, requiring only that the gene of interest is cloned in the correct orientation and is not separated from the plasmid-localized promoter by a transcriptional termination sequence. If all expression signals assigned to a certain gene are recognizable by the host strain, expression is independent from vector signals and identification of the gene as a result of the activity of its product is most likely, provided that no repression occurs.

Many studies have revealed the enormous diversity of mostly unculturable bacterial species in different natural habitats (Torsvik and Øvreås, 2002; Bohannan and Hughes, 2003), which has served as an incentive to apply random expression cloning techniques to environmental DNA for the recovery of novel enzymatic activities. However, little is known about the question to which extent the metagenome can actually be exploited by this strategy, i.e. which fraction of the encoded enzymes can be expressed and, consequently, discovered in a heterologous host such as *E. coli*. In this study, we approached this question by deriving a set of formulas that can be used to calculate the number of clones that is needed for the comprehensive screening of an environmental DNA sample, taking into account the three basic types of heterologous gene expression. Furthermore, a bioinformatics approach was used to predict in which mode the proteins encoded by the genomes of typical soil organisms would be expressed in an *E. coli* host and, consequently, how readily they could be recovered by random expression cloning.

Results and discussion

Statistical analysis

Whether or not a certain gene is discovered by random expression cloning statistically depends on the size of the screened gene bank. Even a gene without any expression signal can be expressed as a fusion protein and, consequently, be recovered, provided that a large enough gene bank (>10^7 clones, see below) is tested. For practical reasons, however, the number of clones that can be prepared and screened is rather limited and typically ranges from 10^4 to 10^6 clones.

The number of clones required to detect an enzymatic activity with a certain probability depends on the expression mode of the corresponding gene. In principle, assaying a single clone that carries a random heterologous DNA fragment can be regarded as a *Bernoulli* experiment with only two possible outcomes: the respective clone exhibits activity or it is inactive. A positive result can be expected with the probability n_{active}/n_{total}, with n_{active} as the number of different active clones that can theoretically be constructed and n_{total} as the number of all possible constructs that can be made from a given DNA sample. As n_{total} is much larger than the typically screened library, n_{active}/n_{total} can be assumed to be constant for all tested clones. Consequently, the number of clones N_P that is required to recover a target gene at least once with the probability P, can be derived from a binomial distribution:

$$N_P = \frac{\ln(1-P)}{\ln\left(1-\frac{n_{active}}{n_{total}}\right)} \quad (1)$$

For gene banks prepared from a single organism, n_{active}/n_{total} is given by $(I\text{-}X)/(c\cdot G)$ as illustrated in Fig. 2, with I as the insert size, X as the size of the gene of

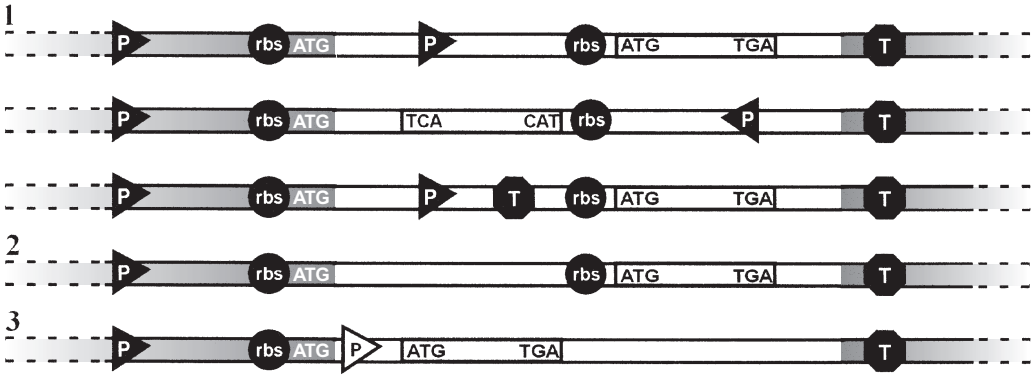

Fig. 1. Different modes of gene expression in a heterologous bacterial host. Vector DNA is shaded in grey, insert DNA is white. P, promoter; T, transcription terminator. (1) Independent gene expression (IND). The insert can be cloned in either direction. However, if a transcriptional termination sequence is located in-between the present promoter and the cds start, the gene can only be expressed as a transcriptional fusion (bottom). (2) Expression based on a transcriptional fusion (TRANSC). (3) Expression dependent on expression signals located on the vector (DEP). Here, a promoter that is active in the host system may be present (white triangle). However, a suitable rbs is lacking, which is why the gene can only be expressed as a fusion protein.

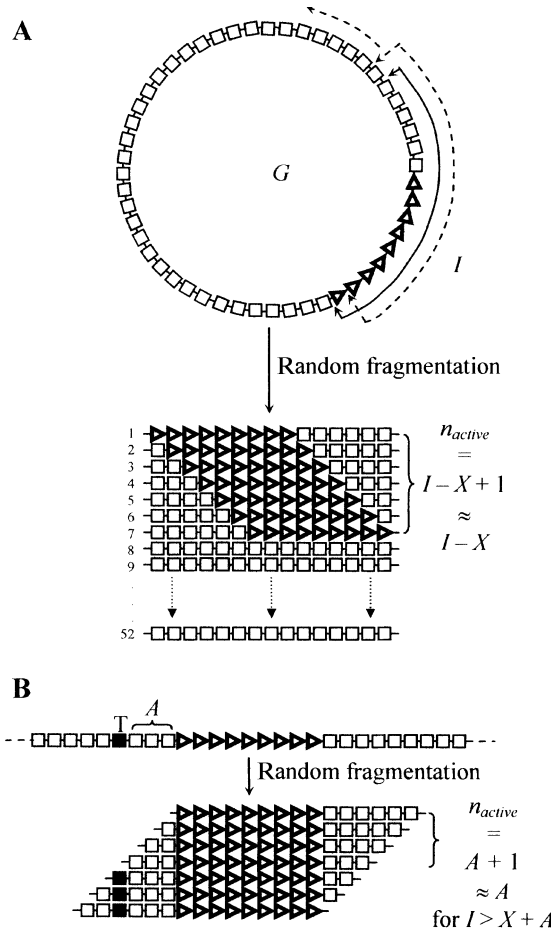

Fig. 2. Schematic overview of the different inserts that can be produced by random fragmentation of a single genome and exemplary calculation of n_{active} and n_{total}. To illustrate the principle, convenient but biologically meaningless values are chosen for the different parameters (genome size $G = 52$ bp, and insert size $I = 15$ bp). Only a single target gene (gene size $X = 9$ bp, including potential promoter and rbs sequences) is present (triangles). Inserts that are randomly prepared can start at any position in the genome, leading to a total of 52 (= G) different possible clones.
A. In this way, I-X inserts with a complete gene can be obtained, leading to a corresponding number of active clones in the case of independent gene expression. Consequently, n_{active}/n_{total} is $(I\text{-}X)/(G \cdot c)$ with $c = 1$. If transcriptional fusions are required, the inserts carrying a complete gene need to be cloned in the right direction behind the vector promoter, yielding active clones in only half of the cases as reflected by $c = 2$.
B. If a transcriptional fusion is needed and $I > A + X$ with A as the distance between the gene start and the closest upstream terminator (T), only A active clones can be formed irrespective of the used insert size. In this case, n_{active}/n_{total} becomes $A/(2 \cdot G)$.

interest and G as the genome size. The correction factor c reflects the different possible expression modes, IND, TRANSC and DEP, as described below. For simplicity, the target enzyme is assumed to be encoded by a single copy gene.

When libraries are prepared from metagenomic DNA, G becomes the average size of genome sequences present in the sample, and z is the number of different genomes (species) comprised, assuming their even distribution in the source DNA:

$$N_P = \frac{\ln(1-P)}{\ln\left(1-\frac{I-X}{G \cdot c \cdot z}\right)} \quad (2)$$

From all genome sequences of prokaryotes available to date at The Institute for Genomic Research (http://www.tigr.org), the average size is 3100 kb, a number that can be used as a rough approximation for G. This value, however, is somewhat biased towards pathogenic organisms, which are predominant in the database and have been found to be considerably smaller than their free-living, environmental counterparts (Horn *et al.*, 2003). An estimate for z could be obtained by analysing different soil and sediment samples by denaturing gradient gel electrophoresis, revealing 25–44 different species (Gabor *et al.*, 2003).

For independent gene expression, c is 1, leading to relatively small numbers of clones that need to be screened to recover a target gene (Fig. 2A). If transcriptional fusions are required, however, c equals 2 as the genes need to be cloned in the correct orientation to the vector promoter and, consequently, N_P increases. For the DEP class of genes, N_P rises even further because here, c = 6 as genes must be specifically cloned into one of the 6 reading frames. It should be noted, however, that these considerations for the TRANSC and DEP genes only hold true when relatively small inserts are used. For larger inserts, the situation becomes more complicated as described below.

It follows from Equation 2 that by using larger insert sizes, N_P decreases, which is obviously a desired effect and especially holds for the IND fraction of genes. If transcription of a given cds, however, needs to be triggered from the vector molecule such as for members of the TRANSC group, also the chance of obtaining non-productive clones rises due to the increased chance of creating constructs that contain a transcription terminator upstream of the target gene. This fact causes N_P to become independent of the chosen insert size if $I > A + X$, A being the average distance between a start codon and its preceding transcription terminator (Fig. 2B):

$$N_P = \frac{\ln(1-P)}{\ln\left(1-\frac{A}{2 \cdot G \cdot z}\right)} \quad (3)$$

Consequently, choosing inserts larger than $A + X$ will not further decrease the number of clones that need to be screened.

If translational fusions are required, the useful insert size is further reduced. Even if a cds is cloned into the right reading frame and no transcription terminator is located upstream of it, two situations can occur that com-

promise the formation of an active gene product: (i) a stop codon is located between the rbs/start codon stretch of the vector and the actual cds start, or (ii) the amino acids fused to the N-terminus of the resulting fusion protein impede activity. The number of tolerated fused amino acids *B* is difficult to determine in general terms although it can be anticipated that it is rather small, resulting in the requirement to use extremely large gene banks according to Equation 4, which can be deduced in a similar way as Equation 3.

$$N_P = \frac{\ln(1-P)}{\ln\left(1-\frac{3\cdot B}{6\cdot G\cdot z}\right)} \text{ for } I > 3\cdot B + X \qquad (4)$$

The GeneClassifier program

As becomes apparent from the statistical considerations described above, the likelihood of discovering environmental genes, i.e. their accessibility by random cloning, is directly coupled to the way they can be expressed in a heterologous host. To estimate the fraction of the metagenome that can be mined using *E. coli* as the expression host, we developed a program called GeneClassifier (Fig. 3).

GeneClassifier locates putative promoter, rbs and transcriptional terminator sequences in complete genome sequences, using matrix searches with *E. coli* consensus sequences and the TransTerm program respectively. The locations of cds within the genome sequences are taken from the annotations. Searches for promoter sequences are performed in the 2500 bp preceding each cds, whereas rbs searches are restricted to the −12 to +3 region. The positions of possible promoters and rbs are determined by searching the genomic sequences with positional weight matrices (PWMs) (Stormo *et al*., 1986) that describe the consensus sequences for those sites and can be obtained from positional frequency matrix (PFMs). Positional frequency matrices of *E. coli* promoters and rbs were taken from Lisser and Margalit (1993) and Schneider and Stephens (1990) respectively (Fig. 4). GeneClassifier scores sequences based on their match to the PWMs using the Perl TFBS modules (Lenhard and Wasserman, 2002). The absolute score of each site (S_{site}) is then converted to a relative score (S_{rel}), given by

$$S_{rel} = \frac{S_{site} - S_{min}}{S_{max} - S_{min}} \qquad (5)$$

in which S_{min} and S_{max} are the scores of the worst and the best possible match to the PWM respectively. All sites that score above a given cut-off value of S_{rel} are considered to be putative promoters or rbs and are assigned to the cds they are preceding. It should be noted that the values of the cut-offs reflect the respective promoter and rbs strength, and are thus directly related to the expression level. When using a sensitive screening assay, weaker promoter and rbs sites can in principle be accepted than when using tests that require high amounts of enzyme activity. The locations of intrinsic (rho-independent) transcription terminators are determined by screening the complete intergenic regions of the genomes with the TransTerm program (http://www.tigr.org/software/) (Ermolaeva *et al*., 2000).

Based on the relative positions of the cds and the (possibly) present upstream expression signals, the transcriptional and translational context of each cds is then determined according to Fig. 1, and the frequencies of the three categories (IND, TRANSC, and DEP) in the analysed genome are calculated. Cds with a preceding rbs and a promoter, and without a transcription terminator in-between the promoter and the cds are classified as IND. Genes with only a predicted rbs as well as genes with a valid rbs and a promoter but with an intervening terminator are assigned to the TRANSC group of genes, whereas genes lacking both the rbs and the promoter are labelled as DEP.

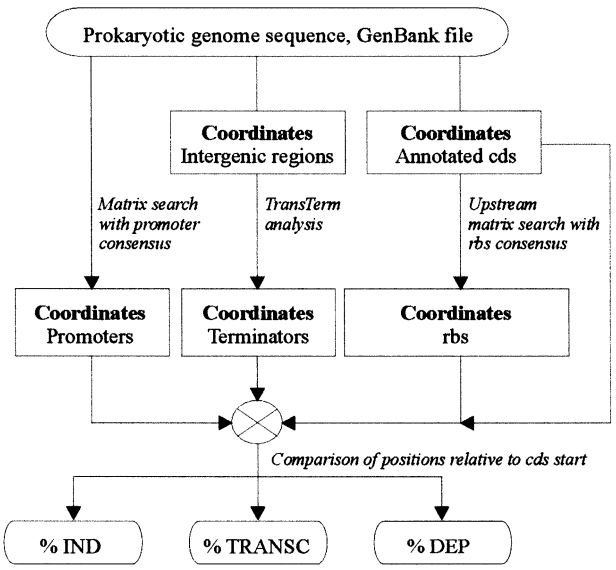

Fig. 3. Architecture of the GeneClassifier program.

Heterologous gene expression in E. coli

A total of 32 completely sequenced prokaryotic genomes (Table 1) were analysed with the GeneClassifier program. As soil and sediment samples are mostly used in metagenomic cloning, we selected genomes of organisms that fall into different classes of bacteria known to be abundant in these habitats: Proteobacteria, Actinobacteria, and low-G/C Gram-positive Eubacteria, i.e. Firmicutes (Handelsman *et al*., 1998). Recently, evidence has been provided

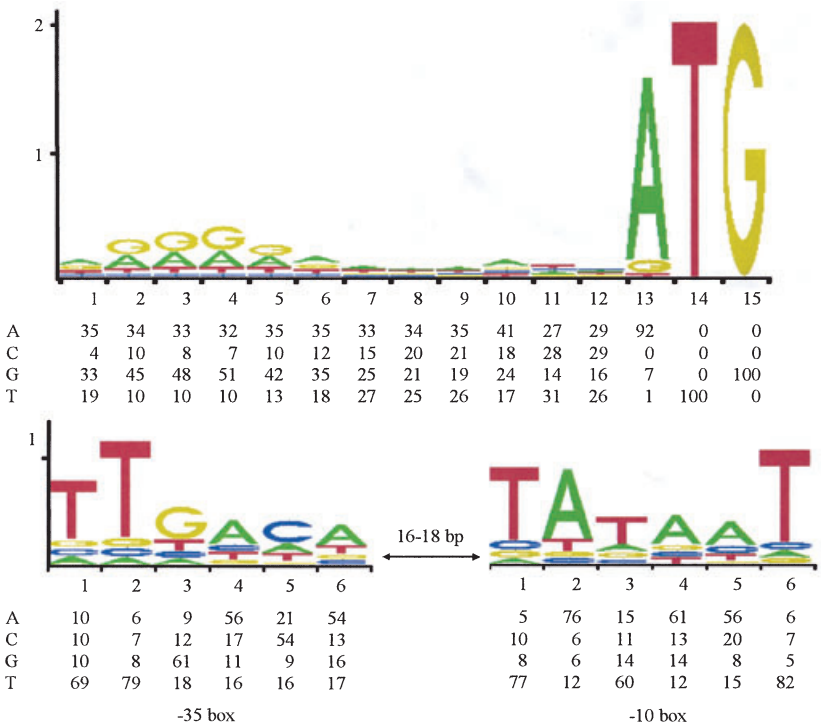

Fig. 4. Sequence logos and positional frequency matrices (PFMs) of the *E. coli* rbs (top, Schneider and Stephens, 1990) and promoter (bottom, Lisser and Margalit, 1993) consensus sequences used in this study. Tabulated values are the frequencies (in percentage) of the four nucleotides A, C, G, T.

Table 1. Genomes analysed in this study. Average distances between the closest terminator upstream of each cds and the start codon (*A*) were calculated with the GeneClassifier program.

Taxonomic group	Species (circular chromosome length [Mb])	Accession no.	GC-content [%]	*A* [kb]
Euryarchaeota	*Archaeoglobus fulgidus* (2.2)	AE000782	48.5	489
	Methanobacterium thermoautotrophicum (1.8)	AE000666	49.5	190
	Methanococcus jannaschii (1.7)	L77117	31.3	40
	Pyrococcus abyssi (1.8)	AL096836	44.6	354
	Thermoplasma acidophilum (1.6)	AL139299	45.9	111
	Methanopyrus kandleri (1.7)	AE009439	61.1	118
	Methanosarcina mazei (4.1)	AE008384	41.4	36
Crenarchaeota	*Aeropyrum pernix* (1.7)	BA000002	56.2	81
	Pyrobaculum aerophilum (2.2)	AE009441	51.3	633
	Sulfolobus solfataricus (3.0)	AE006641	35.7	287
Firmicutes (low-G/C Gram-positive)	*Clostridium acetobutylicum* (3.9)	AE001437	30.9	5
	Listeria monocytogenes (3.0)	AL591824	37.3	4
	Staphylococcus aureus (2.8)	BA000018	32.7	5
	Streptococcus pneumoniae (2.0)	AE007317	39.6	4
	Bacillus subtilis (4.2)	AL009126	43.5	4
Actinobacteria (high-G/C Gram-positive)	*Streptomyces coelicolor* (8.7)	AL645882	72.1	49
	Bifidobacterium longum (2.2)	AE014295	60.0	7
	Corynebacterium efficiens (3.1)	BA000035	53.7	11
	Mycobacterium leprae (3.3)	AL450380	57.7	267
Proteobacteria (Gram-negative)	*Brucella melitensis* (2.1)	AE008917/8	57.0	11
	Rickettsia prowazekii (1.3)	AJ235269	29.1	155
	Agrobacterium tumefaciens (2.8)	AE007869	59.3	8
	Caulobacter crescentus (4.0)	AE005673	67.1	11
	Neisseria meningitidis (2.3)	AE002098	51.4	3
	Escherichia coli (4.6)	U00096	50.7	8
	Pseudomonas aeruginosa (6.3)	AE004091	66.4	16
	Yersinia pestis (4.7)	AL590842	47.5	8
	Haemophilus influenzae (1.8)	L42023	38.0	4
	Vibrio cholerae (3.0)	AE003852/3	47.6	5
	Xylella fastidiosa (2.7)	AE003849	52.6	54
	Heliobacter pylori (1.7)	AE000511	38.8	65

that also Archaea, organisms usually assigned to extreme environments, populate these habitats (Bintrim *et al*., 1997) and DNA samples extracted from different soil types and sediment were found to contain up to 17% of archaeal genomic DNA (Gabor *et al*., 2003). Therefore, also 10 archaeal representatives were included in the expression analysis.

The choice of the cut-off values for the promoter and rbs searches (see above) can have a significant impact on the outcome of the GeneClassifier analysis. Results for the five taxonomic groups of organisms at different combinations of threshold settings are shown in Fig. 5. It turned out that the fraction of genes in the IND class decreases with increasing cut-off values for both the rbs and the promoter sites, whereas the fraction of genes in the TRANSCR class is mainly determined by the respective promoter cut-off. The curves for the different phyla, however, do not cross each other, except in the extreme high-rbs/high-promoter cut-off region of the TRANSC category, which indicates that the differences between taxonomic groups are found for a wide range of biologically meaningful parameter settings.

For the sake of clearness, GeneClassifier results are shown for a single cut-off combination (0.75 for both the promoter and the rbs) in Fig. 6. Firmicutes were predicted to have the largest fraction of independently expressible genes (73%), which is agreement with experimental observations of Handelsman *et al*. (1998) who found that more than 50% of the traits of *Bacillus cereus* that they tested were readily expressed in *E. coli*, presumably from their own expression signals. Actinobacteria, in contrast, only contained 7% of IND genes. Surprisingly, the phylum of Proteobacteria, containing *E. coli* itself, did not constitute the group with the largest IND fraction. This alleged paradox can be partially explained by the different GC-contents of the analysed taxonomic groups (Table 1). Firmicutes have a relatively low GC-content compared with Actinobacteria (37% versus 61% in average), which statistically leads to a higher chance of matches to the AT-rich promoter consensus in the first group of organisms. Corresponding to their GC-contents of 47 and 51%, respectively, Archaea and Proteobacteria ranked at an intermediate position with respect to their IND genes. The apparent presence of many promoters in Archaea, however, is not only explained by statistical reasons. It is known that archaeal promoter sequences resemble eukaryotic promoters in structure and function, but also contain a TTTAWATA (W = A or T) motif about 20 bp upstream of a less important but AT-rich initiator element (Brown *et al*., 1989). These two sequence stretches closely match the *E. coli* – 10 and –35 promoter boxes respectively.

Interestingly, average distances between cds starts and preceding transcription terminators, *A*, were very large for

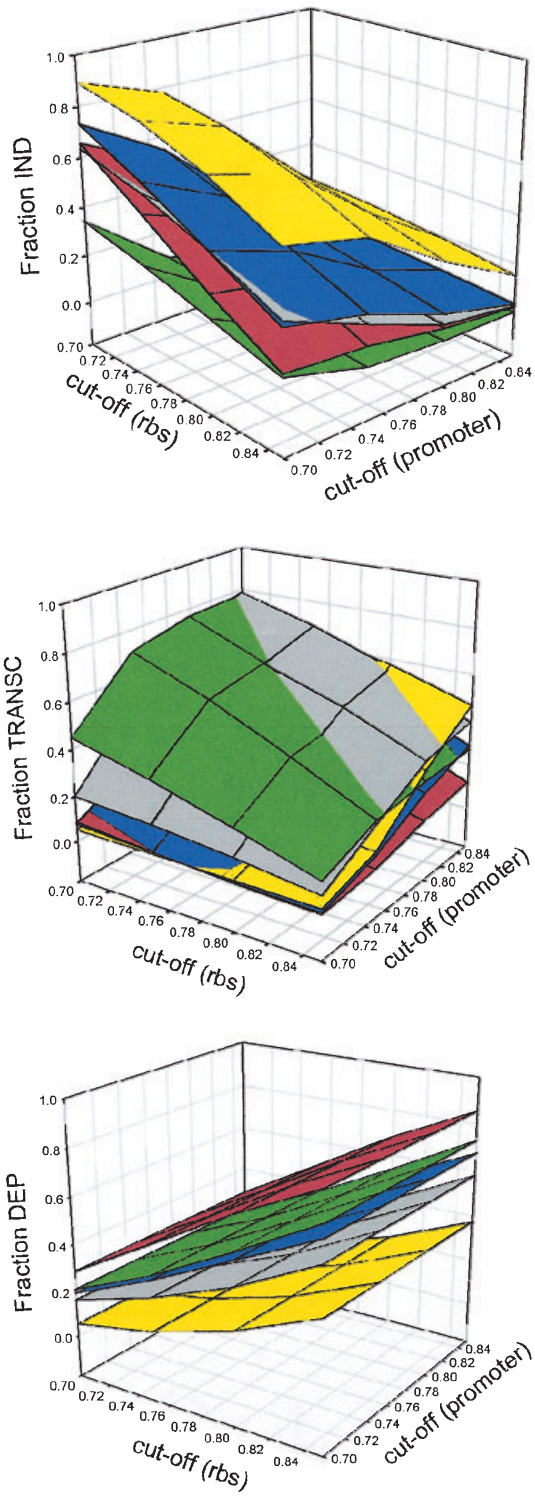

Fig. 5. Dependence of GeneClassifier results on the chosen cut-off values for promoter and rbs searches. Fraction of IND genes (top); fraction of TRANSC genes (middle); fraction of DEP genes (bottom). Euryarchaea ■; Crenarchaea ■; Firmicutes ■; Actinobacteria ■; Proteobacteria ■.

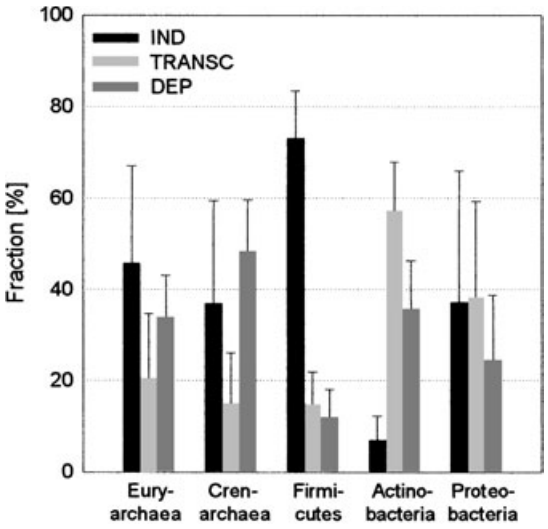

Fig. 6. GeneClassifier results for the different taxonomic groups analysed. For both promoter and rbs searches, a threshold of 0.75 was chosen.

all archaeal genomes analysed by GeneClassifier, ranging clearly above 100 kb for most genomes (Table 1). Apparently, intrinsic transcription termination is not a common mechanism in this group of organisms, which is in agreement with the findings of others (Washio *et al.*, 1998). For expression cloning of the TRANSC fraction of archaeal genes, the absence of transcription terminators is, in fact, a convenient feature as there is no disturbance of transcriptional fusions, and large insert sizes can be used to effectively decrease the number of clones that needs to be screened (see previous section).

In eubacterial genomes, in contrast, *A* was found to be 15 kb in average, excluding the *M. leprae* and *R. prowazekii* genomes, for which unusually low numbers of transcription terminators were predicted. As can be seen in Fig. 6, the TRANSC category constitutes the largest fraction of genes of Actinobacteria and Proteobacteria, and still comprised 15% of the genes of Firmicutes. For the cloning of this large number of genes, a small insert size of around 15 kb thus appears optimal, taking into account the increasing experimental expenditure when working with large DNA fragments and the higher expression levels that can usually be reached in high copy number vectors that are suitable for small insert sizes. It is obvious that these considerations only hold when targeting single genes or small operons. Especially if screening attempts to discover novel natural products, which often require the expression of large biosynthetic gene clusters, large insert vectors such as bacterial artificial chromosomes (BACs) remain the most appropriate vector systems (MacNeil *et al.*, 2001).

With 34% to 48%, Archaea and Actinobacteria contained the largest fractions of genes depending on a translational fusion to the vector molecule and therefore most of their DNA is virtually inaccessible by conventional random expression cloning. As expected, genes of Firmicutes and Proteobacteria were identified to be most readily expressed in *E. coli*.

Conclusion

This study constitutes a quantitative approach to the question to which extent the metagenome can be exploited by the current random expression cloning techniques. The probability of detecting a certain enzyme activity is directly correlated with the expression mode of the respective cds in a heterologous host. As summarized in Fig. 7, genes that are preceded by expression signals that are functional in *E. coli* can be recovered by screening a relatively small number of clones, a number that exponentially decreases when using larger insert sizes. About 40% of the genes of all genomes analysed in this study were predicted to be readily expressible in this way, with strong variations between different groups of organisms (7–73%). The expression of the majority of genes, in contrast, was found to be dependent on expression signals located on the cloning vector. One-third requires transcription to be triggered from the vector promoter. To recover genes from the TRANSC category, significantly bigger gene banks need to be constructed than for the IND fraction, particularly when working with large inserts. In fact, using inserts >15 kb does not seem to be useful given the abundance of transcription terminators that may interfere with the formation of a complete transcript. Another 30% of all analysed genes, again with strong deviations between different phyla, can only be expressed as fusion proteins in *E. coli* due to the lack of suitable expression signals. This demands for gene libraries that comprise more than 10 millions of clones irrespective of

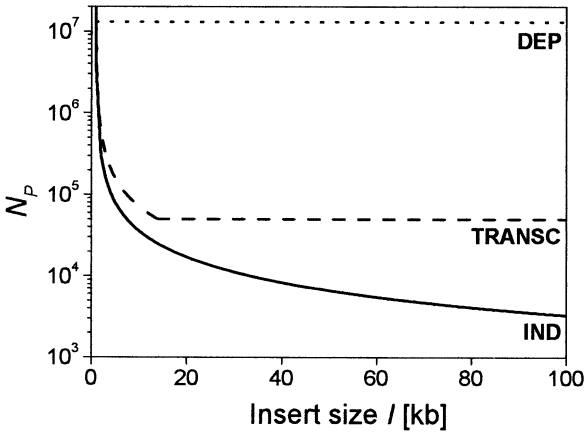

Fig. 7. Dependence of N_P on the chosen insert size for the three different expression modes. Curves were calculated with Eqs 2–4 and $P = 0.9$, $X = 0.9$, $G = 3200$ kb, $z = 44$, $A = 13$ kb, and $B = 50$

the used insert size (Equation 4, Fig. 7), which is not feasible in the scope of most projects. Consequently, genes falling into the DEP category can be regarded to be virtually inaccessible by random expression cloning.

In view of the fact that enzyme activity not only requires protein expression but also proper folding, incorporation into the cell membrane or secretion, and in some cases the presence of specific cofactors or modifying enzymes, the presented estimate of genes that can be detected in functional screenings is relatively optimistic, constituting an upper limit rather than an absolute number. Compared with expression mechanisms, however, post-translational processes are much more complex and diverse, making predictions on their occurence in heterologous hosts extremely difficult. With growing insight into these processes, however, the presented analysis may be refined to allow more precise predictions.

The amount of new proteins to be discovered by random expression cloning thus appears to be not as gigantic as originally thought, and from the total estimated enzymatic diversity of about 10^{13} distinct functional sequences (Burton *et al*., 2002) only a part is expected to be accessible in *E. coli*. Alternative hosts, however, particularly from taxonomic groups that contain only few expression signals functional in *E. coli* (e.g. *Streptomyces* sp. from the group of Actinobacteria), may significantly broaden the range of exploitable genes.

Experimental procedures

The GeneClassifier program as outlined in the *Results* section was written as a pipeline of Perl scripts using Bioperl modules (Stajich *et al*., 2002). Positional frequency matrices for *E. coli* promoters and rbs were obtained from literature (Schneider and Stephens, 1990; Lisser and Margalit, 1993) and are shown in Fig. 4. Intrinsic transcription terminators were predicted with the TransTerm program (http://www.tigr.org/software/) (Ermolaeva *et al*., 2000). The annotated sequence files of the genomes used in this report were downloaded from the GenBank database at the National Center for Biotechnology Information (NCBI, ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria).

References

Bintrim, S.B., Donohue, T.J., Handelsman, J., Roberts, G.P., and Goodman, R.M. (1997) Molecular phylogeny of Archaea from soil. *Proc Natl Acad Sci USA* **94:** 227–282.

Bohannan, B.J.M., and Hughes, J. (2003) New approaches to analysing microbial biodiversity data. *Curr Opin Microbiol* **6:** 282–287.

Brown, J.W., Daniels, C.J., and Reeve, J.N. (1989) Gene structure, organization, and expression in archaeabacteria. *CRC Crit Rev Microbiol* **16:** 287–337.

Burton, S.G., Cowan, D.A., and Woodley, J.M. (2002) The search for the ideal biocatalyst. *Nature Biotechnol* **20:** 37–45.

Cowan, D.A. (2000) Microbial genomes – the untapped resource. *Trends Biotechnol* **18:** 14–16.

Ermolaeva, M.D., Khalak, H.G., White, O., Smith, H.O., and Salzberg, S.L. (2000) Prediction of transcription terminators in bacterial genomes. *J Mol Biol* **301:** 27–33.

Gabor, E.M., de Vries, E.J., and Janssen, D.B. (2003) Efficient recovery of environmental DNA for expression cloning by indirect extraction methods. *FEMS Microbiol Ecol* **44:** 153–163.

Gold, L., Pribnow, D., Schneider, T., Shinedling, S., Singer, B., and Stormo, G. (1981) Translational initiation in prokaryotes. *Annu Rev Microbiol* **35:** 365–403.

Handelsman, J., Rondon, M.R., Brady, S.F., Clardy, J., and Goodman, R.M. (1998) Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol* **5:** R245–R249.

Horn, M., Collingro, A., Schmitz-Esser, S., Purkhold, U., Beier, C., Fartmann, B., Brandt, P., *et al.* (2003) EDGE – The environmental chlamydiae genome project. Darmstadt, Germany: Metagenomics.

Lenhard, B., and Wasserman, W.W. (2002) TFBS: Computational framework for transcription factor binding site analysis. *Bioinformatics* **18:** 1135–1136.

Lisser, S., and Margalit, H. (1993) Compilation of *E. coli* mRNA promoter sequences. *Nucleic Acids Res* **21:** 1507–1516.

Lorenz, P., and Schleper, C. (2002) Metagenome – a challenging source of enzyme discovery. *J Mol Catal B-Enzym* **19–20:** 13–19.

MacNeil, I.A., Tiong, C.L., Minor, C., August, P.R., Grossman, T.H., Loiacono, K.A., Lynch, B.A., *et al.* (2001) Expression and isolation of antimicrobial small molecules from soil DNA libraries. *J Mol Biotechnol* **3:** 301–308.

Schneider, T.D., and Stephens, R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res* **18:** 6097–6100.

Staden, R. (1983) Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Res* **12:** 509–519.

Stajich, J.E., Block, D., Boulez, K., Brenner, S.E., Chervitz, S.A., Dagdigian, C., *et al.* (2002) The Bioperl toolkit: perl modules for the life sciences. *Genome Res* **12:** 1611–1618.

Stormo, G.D., Schneider, T.D., and Gold, L. (1986) Quantitative analysis of the relationship between sequence and functional activity. *Nucleic Acids Res* **14:** 6661–6679.

Torsvik, V., and Øvreås, L. (2002) Microbial diversity and function in soil: from genes to ecosystems. *Curr Opin Microbiol* **5:** 240–245.

Washio, T., Sasayama, J., and Tomita, M. (1998) Analysis of complete genomes suggests that many prokaryotes do not rely on hairpin formation in transcription termination. *Nucleic Acids Res* **26:** 5456–5463.